

said plurality of entities operating in parallel on work partitions assigned to them to perform said operation.

21. The method of Claim 20 wherein:

the method further includes the step of receiving a query that requires the operation;

and

the step of dividing the operation is performed in response to processing the query.

22. The method of Claim 20 wherein the step of assigning work partitions includes:

assigning said at least one entity a first work partition from said set of work

partitions; and

after said at least one entity has completed operating on said first work partition,

assigning said at least one entity a second work partition from said set of work partitions.

23. The method of Claim 20 wherein said plurality of entities are a plurality of processes.

24. The method of Claim 23 wherein said plurality of processes reside within a single database system.

25. The method of Claim 20 wherein the step of dividing an operation into a set of work partitions includes generating a plurality of query fragments for at least a portion of said operation.

26. The method of Claim 25 wherein:

the operation corresponds to at least a portion of a query; and

the plurality of query fragments are generated based on said query.

27. The method of Claim 20 the step of assigning work partitions is performed by

assigning the work partitions in a sequence based at least in part on sizes associated with the
work partitions.

28. The method of Claim 27 wherein the step of assigning the work partitions in a

sequence is performed by assigning relatively larger work partitions before assigning
relatively smaller work partitions.

29. The method of Claim 22 wherein:

the operation is specified in a query that corresponds to a hierarchy of operations; and

the step of assigning said at least one entity a second work partition includes

determining whether there are any unassigned work partitions from a first

level in the hierarchy to which said first work partition belonged; and

if there are no unassigned work partitions from the first level in the hierarchy,

then selecting said second work partition from a level in said hierarchy

that is two levels below said first level in said hierarchy.

30. The method of Claim 20 wherein:

a'
the method includes the step of generating a serial execution plan for operations in a database management system (DBMS) running on a computer system;
the method includes the step of generating a parallelized execution plan for said serial execution plan, said parallelized execution plan including first and second operations;
the step of dividing an operation is performed by dividing said second operation;
the plurality of entities includes one or more slave processes operating on a plurality of data partitions, the quantity of said data partitions being greater than the quantity of said slave processes;
executing said parallelized execution plan when a plurality of parallel resources of said computer system are available; and
executing said serial execution plan when said plurality of resources are not available.

31. The method of claim 30 wherein said step of generating a parallelized execution plan includes the steps of:
identifying one or more segments of said serial execution plan that can be parallelized; and
identifying partitioning requirements of said one or more segments

32. The method of claim 30 wherein said step of generating a parallelized execution plan is based on a specification of parallelism in a statement specifying one of said operations.

33. The method of Claim 20 further comprising the steps of:

a

generating an execution plan for said operation;
examining said execution plan from bottom up;
identifying a parallelized portion of said execution plan, said parallelized portion can
be processed in parallel, said parallelized portion including first and second
operations, said first and second operations being executable in parallel;
wherein the step of dividing the operation is performed by dividing said second
operation;
wherein the plurality of entities includes one or more slave processes operating on a
plurality of data partitions, the quantity of said data partitions being greater
than the quantity of said slave processes;
identifying some serial portion of said execution plan, said serial portion can be
processed in serial;
allocating a central scheduler between said parallelized portion and said serial
portion.

34. The method of Claim 33 further including the steps of:

identifying a first data flow requirement for a first portion of said execution plan said
first data flow requirement corresponding to a partitioning of a data flow
required by said first portion;
identifying a second data flow requirement for a second portion of said execution
plan said second data flow requirement corresponding by said second portion;
and

allocating a data flow director between said first portion and said second portion

when said first data flow requirement is not compatible with said second data flow requirement said data flow director repartitioning a data flow of said first portion to be compatible with said second data flow requirement

a!
35. The method of Claim 20 further comprising the steps of:

generating an execution plan to execute database management system (DBMS)

operations in parallel, said execution plan including first and second operations;

wherein the step of dividing said operation is performed by dividing said second operation;

initiating an operation coordinator in a computer system to coordinate execution of said execution plan;

initiating, by said operation coordinator, a first set of slaves operating on a plurality of data partitions to produce data, the quantity of said data partitions being greater than the quantity of said first set of slave processes;

initiating, as said plurality of entities, by said operation coordinator, a second set of slaves to consume data; and

directing said second set of slaves to produce data and said first set of slaves to consume data when said first set of slaves finishes producing data.

36. The method of claim 35 wherein said execution plan is comprised of operator nodes and said operator nodes are linked together to form execution sets

37. The method of Claim 20 further comprising the steps of:

generating an execution plan to execute said operations in parallel, said execution plan including first and second operations;

wherein the step of dividing said operation includes dividing said first operation;

initiating a data flow scheduler in said computer system to coordinate data flow;

initiating, as said plurality of entities, by said data flow scheduler, producer slaves operating on a plurality of data partitions to produce a first data production;

initiating, by said data flow scheduler, consumer slaves to consume said first data production;

transmitting a ready message to said data flow scheduler when said producer slaves become ready to produce data;

transmitting a completion message to said data flow scheduler when said first data production is completed;

generating, by said data flow scheduler, in response to said completion message, an identification of a plurality of said consumer slaves that did not receive data in said first data production, said generating step using information derived from said ready message;

examining, by said producer slaves, said identification during a subsequent data production; and

reducing said subsequent data production such that said subsequent data production does not produce data for said plurality of said consumer slaves

a' 38. A method for processing a query, the method comprising the steps of:
receiving a statement that specifies at least (a) an operation and (b) a degree of
parallelism to use in performing the operation;
dividing the operation into a set of work partitions;
performing a determination of how many entities to use to perform said operation
based, at least in part, on the degree of parallelism specified in said statement;
assigning work partitions from said set of work partitions to a plurality of entities
based on said determination; and
said plurality of entities operating in parallel on work partitions assigned to them to
perform said operation.

39. The method of Claim 38 wherein:
the query requires a plurality of operations; and
the statement specifies said degree of parallelism for a subset of the plurality of
operations required by the query.

40. The method of Claim 39 wherein the degree of parallelism specified by the query
indicates that no amount of parallelism is to be used during execution of a particular portion
of the query.

41. The method of Claim 38 wherein the degree of parallelism specified by the query
indicates a maximum amount of parallelism to use during execution of said operation.

a¹

42. A method of processing a query, the method comprising the steps of:
dividing an operation required by said query into a set of work partitions by
generating a set of query fragments;
incorporating hints into at least some of said query fragments, wherein the hint
associated with a given query fragment indicates how to perform the work
partition associated with said given query fragment;
assigning query fragments from said set of query fragments to a plurality of entities;
and
said plurality of entities operating in parallel on query fragments assigned to them to
perform said operation, wherein entities working on a query fragment
associated with a hint perform the work partition associated with said query
fragment in a manner dictated by said hint.

43. The method of Claim 42 wherein the step of incorporating hints includes
incorporating hints that dictate the operation of a table scan.

Sub DS
44. ~~The method of Claim 43 wherein the step of incorporating hints that dictate the~~
~~operation of a table scan includes incorporating hints that rowed partitioning is to be used~~
~~during the table scan.~~

45. The method of Claim 42 wherein the step of incorporating hints includes
incorporating hints that specify performance of a full table scan.

46. The method of Claim 42 wherein the step of incorporating hints includes incorporating hints that specify using a particular type of join.

a' 47. The method of Claim 46 wherein the step of incorporating hints that specify using a particular type of join includes incorporating hints that specify using a sort/merge join.

48. The method of Claim 46 wherein the step of incorporating hints that specify using a particular type of join includes incorporating hints that specify using a nested loop join.

49. A method of processing a query, the method comprising the steps of:
determining a hierarchy of operations associated with a query;
dividing a first operation required by said query into a first set of work partitions;
dividing a second operation required by said query into a second set of work partitions, wherein said second operation immediately follows said first operation in said hierarchy;
dividing a third operation required by said query into a third set of work partitions, wherein said third operation immediately follows said second operation in said hierarchy;
assigning work partitions from said first set of work partitions to a first plurality of entities;
said first plurality of entities operating in parallel on work partitions assigned to them from said first set of work partitions to perform said first operation;

a' assigning work partitions from said second set of work partitions to a second plurality of entities, wherein said second plurality of entities are different entities than said first plurality of entities; and
said second plurality of entities operating in parallel on work partitions assigned to them from said second set of work partitions to perform said second operation;

assigning work partitions from said third set of work partitions to said first plurality of entities; and
said first plurality of entities operating in parallel on work partitions assigned to them from said third set of work partitions to perform said third operation.

50. The method of Claim 49 further comprising performing the following steps when a given entity in said first set of entities finishes performing a work partition from said first set of work partitions:

determining whether there are any unassigned work partitions from said first set of work partitions; and
if there are no unassigned work partitions from said first set of work partitions, then assigning the given entity a work partition selected from said third set of work partitions; and
if there are unassigned work partitions from said first set of work partitions, then assigning the given entity a work partition selected from said first set of work partitions.

51. The method of Claim 49 wherein the hierarchy includes odd levels and even levels, and the method further comprises the steps of assigning work partitions from odd levels to said first plurality of entities and work partitions from even levels to said second plurality of entities.

52. The method of Claim 49 wherein performing work partitions in said first set of work partitions causes said first set of entities produce output consumed by said second plurality of entities, and performing work partitions in said third set of work partitions causes said first set of entities to consume output produced by said second plurality of entities.

53. A method of processing a query, the method comprising the steps of:
determining that, to execute said query, output from a plurality of producer operations
is to be supplied to a consumer operation;
wherein a first set of entities is responsible for executing a first producer operation of
said plurality of producer operations;
wherein a second set of entities is responsible for executing a second producer
operation of said plurality of producer operations;
wherein a third set of entities is responsible for executing said consumer operation;
during execution of said query, performing the steps of
determining whether any entity in said first set of entities produced output for
a particular entity in said third set of entities; and
if no entity in said first set of entities produced output for said particular entity
in said third set of entities, then communicating to at least one entity in

said second set of entities that is responsible for supplying output to
said particular entity that said at least one entity need not produce
output for said particular entity.

a'
54. The method of Claim 53 wherein the step of determining whether any entity in said
first set of entities produced output for a particular entity in said third set of entities is
performed by

the particular entity monitoring whether it received any output from any entity in said
first set of entities;

the particular entity generating data that indicates whether it received any output from
any entity in said first set of entities; and

determining whether any entity in said first set of entities produced output for said
particular entity based on said data.

55. The method of Claim 53 further comprising the steps of:

maintaining a bit vector, wherein each entity in said third set of entities corresponds
to a bit in the bit vector;

when all entities in said first set of entities that produce output for said particular
entity have completed their portion of said first producer operation, setting the
bit, in the bit vector, that corresponds to said particular entity if said particular
entity received no output from said first producer operation; and

wherein the step of determining whether any entity in said first set of entities
produced output for said particular entity includes the step of inspecting said

bit vector to determine whether any entity in said first set of entities produced output for said particular entity.

a' 36. A computer-readable medium carrying instructions for parallelizing an operation, the instructions including instructions for performing the steps of:

dividing the operation into a set of work partitions;

assigning work partitions from said set of work partitions to a plurality of entities,

wherein at least one entity of said plurality of entities is assigned a plurality of

work partitions from said set of work partitions; and

said plurality of entities operating in parallel on work partitions assigned to them to

perform said operation.

57. The computer-readable medium of Claim 56 wherein:

the instructions further include instructions for performing the step of receiving a

query that requires the operation; and

the step of dividing the operation is performed in response to processing the query.

58. The computer-readable medium of Claim 56 wherein the step of assigning work partitions includes:

assigning said at least one entity a first work partition from said set of work

partitions; and

after said at least one entity has completed operating on said first work partition,
assigning said at least one entity a second work partition from said set of work
partitions.

a'
59. The computer-readable medium of Claim 56 wherein said plurality of entities are a
plurality of processes.

60. The computer-readable medium of Claim 59 wherein said plurality of processes
reside within a single database system.

61. The computer-readable medium of Claim 56 wherein the step of dividing an
operation into a set of work partitions includes generating a plurality of query fragments for
at least a portion of said operation.

62. The computer-readable medium of Claim 61 wherein:
the operation corresponds to at least a portion of a query; and
the plurality of query fragments are generated based on said query.

Sub D6
7
63. The computer-readable medium of Claim 56 wherein the step of assigning work
partitions is performed by assigning the work partitions in a sequence based at least in part on
sizes associated with the work partitions.

64. The computer-readable medium of Claim 63 wherein the step of assigning the work partitions in a sequence is performed by assigning relatively larger work partitions before assigning relatively smaller work partitions.

a'
Sub
B3

65. The computer-readable medium of Claim 58 wherein:
the operation is specified in a query that corresponds to a hierarchy of operations; and
the step of assigning said at least one entity a second work partition includes
determining whether there are any unassigned work partitions from a first
level in the hierarchy to which said first work partition belonged; and
if there are no unassigned work partitions from the first level in the hierarchy,
then selecting said second work partition from a level in said hierarchy
that is two levels below said first level in said hierarchy.

Sub
D7

66. The computer-readable medium of Claim 56 wherein:
the instructions include instructions for performing the step of generating a serial
execution plan for operations in a database management system (DBMS)
running on a computer system;
the instructions include instructions for performing the step of generating a
parallelized execution plan for said serial execution plan, said parallelized
execution plan including first and second operations;
the step of dividing an operation is performed by dividing said second operation;

the plurality of entities includes one or more slave processes operating on a plurality of data partitions, the quantity of said data partitions being greater than the quantity of said slave processes;

the instructions include instructions for performing the step of executing said parallelized execution plan when a plurality of parallel resources of said computer system are available; and

the instructions include instructions for performing the step of executing said serial execution plan when said plurality of resources are not available.

67. The computer-readable medium of claim 66 wherein said step of generating a parallelized execution plan includes the steps of:

identifying one or more segments of said serial execution plan that can be parallelized; and

identifying partitioning requirements of said one or more segments

68. The computer-readable medium of claim 66 wherein said step of generating a parallelized execution plan is based on a specification of parallelism in a statement specifying one of said operations

69. The computer-readable medium of Claim 56 further comprising instructions for performing the steps of:

generating an execution plan for said operation;

examining said execution plan from bottom up;

a'

identifying a parallelized portion of said execution plan, said parallelized portion can
be processed in parallel, said parallelized portion including first and second
operations, said first and second operations being executable in parallel;
wherein the step of dividing the operation is performed by dividing said second
operation;
wherein the plurality of entities includes one or more slave processes operating on a
plurality of data partitions, the quantity of said data partitions being greater
than the quantity of said slave processes;
identifying some serial portion of said execution plan, said serial portion can be
processed in serial;
allocating a central scheduler between said parallelized portion and said serial portion

70. The computer-readable medium of Claim 69 further including instructions for
performing the steps of:

identifying a first data flow requirement for a first portion of said execution plan said
first data flow requirement corresponding to a partitioning of a data flow
required by said first portion;
identifying a second data flow requirement for a second portion of said execution
plan said second data flow requirement corresponding by said second portion;
and
allocating a data flow director between said first portion and said second portion
when said first data flow requirement is not compatible with said second data

flow requirement said data flow director repartitioning a data flow of said first portion to be compatible with said second data flow requirement

a1
71. The computer-readable medium of Claim 56 further comprising instructions for performing the steps of:

generating an execution plan to execute database management system (DBMS) operations in parallel, said execution plan including first and second operations;

wherein the step of dividing said operation is performed by dividing said second operation;

initiating an operation coordinator in a computer system to coordinate execution of said execution plan;

initiating, by said operation coordinator, a first set of slaves operating on a plurality of data partitions to produce data, the quantity of said data partitions being greater than the quantity of said first set of slave processes;

initiating, as said plurality of entities, by said operation coordinator, a second set of slaves to consume data; and

directing said second set of slaves to produce data and said first set of slaves to consume data when said first set of slaves finishes producing data.

72. The computer-readable medium of claim 71 wherein said execution plan is comprised of operator nodes and said operator nodes are linked together to form execution sets

73. The computer-readable medium of Claim 56 further comprising instructions for performing the steps of:

generating an execution plan to execute said operations in parallel, said execution plan including first and second operations;

wherein the step of dividing said operation includes dividing said first operation;

initiating a data flow scheduler in said computer system to coordinate data flow;

initiating, as said plurality of entities, by said data flow scheduler, producer slaves operating on a plurality of data partitions to produce a first data production;

initiating, by said data flow scheduler, consumer slaves to consume said first data production;

transmitting a ready message to said data flow scheduler when said producer slaves become ready to produce data;

transmitting a completion message to said data flow scheduler when said first data production is completed;

generating, by said data flow scheduler, in response to said completion message, an identification of a plurality of said consumer slaves that did not receive data in said first data production, said generating step using information derived from said ready message;

examining, by said producer slaves, said identification during a subsequent data production; and

reducing said subsequent data production such that said subsequent data production does not produce data for said plurality of said consumer slaves

74. A computer-readable medium storing instructions for processing a query, the instructions including instructions for performing the steps of:

receiving a statement that specifies at least (a) an operation and (b) a degree of

parallelism to use in performing the operation;

dividing the operation into a set of work partitions;

performing a determination of how many entities to use to perform said operation

based, at least in part, on the degree of parallelism specified in said statement;

assigning work partitions from said set of work partitions to a plurality of entities

based on said determination; and

said plurality of entities operating in parallel on work partitions assigned to them to

perform said operation.

75. The computer-readable medium of Claim 74 wherein:

the query requires a plurality of operations; and

the statement specifies said degree of parallelism for a subset of the plurality of

operations required by the query.

76. The computer-readable medium of Claim 75 wherein the degree of parallelism specified by the query indicates that no amount of parallelism is to be used during execution of a particular portion of the query.

77. The computer-readable medium of Claim 74 wherein the degree of parallelism specified by the query indicates a maximum amount of parallelism to use during execution of said operation.

a¹
78. A computer-readable medium carrying instructions for processing a query, the instructions including instructions for performing the steps of:

dividing an operation required by said query into a set of work partitions by

generating a set of query fragments;

incorporating hints into at least some of said query fragments, wherein the hint

associated with a given query fragment indicates how to perform the work

partition associated with said given query fragment;

assigning query fragments from said set of query fragments to a plurality of entities;

and

said plurality of entities operating in parallel on query fragments assigned to them to

perform said operation, wherein entities working on a query fragment

associated with a hint perform the work partition associated with said query

fragment in a manner dictated by said hint.

79. The computer-readable medium of Claim 78 wherein the step of incorporating hints includes incorporating hints that dictate the operation of a table scan.

Sub
B
A'

~~80. The computer-readable medium of Claim 79 wherein the step of incorporating hints that dictate the operation of a table scan includes incorporating hints that rowed partitioning is to be used during the table scan.~~

81. The computer-readable medium of Claim 78 wherein the step of incorporating hints includes incorporating hints that specify performance of a full table scan.

82. The computer-readable medium of Claim 78 wherein the step of incorporating hints includes incorporating hints that specify using a particular type of join.

83. The computer-readable medium of Claim 82 wherein the step of incorporating hints that specify using a particular type of join includes incorporating hints that specify using a sort/merge join.

84. The computer-readable medium of Claim 82 wherein the step of incorporating hints that specify using a particular type of join includes incorporating hints that specify using a nested loop join.

85. A computer-readable medium carrying instructions for processing a query, the instructions including instructions for performing the steps of:

determining a hierarchy of operations associated with a query;

dividing a first operation required by said query into a first set of work partitions;

dividing a second operation required by said query into a second set of work

partitions, wherein said second operation immediately follows said first
operation in said hierarchy;

dividing a third operation required by said query into a third set of work partitions,
wherein said third operation immediately follows said second operation in
said hierarchy;

assigning work partitions from said first set of work partitions to a first plurality of
entities;

said first plurality of entities operating in parallel on work partitions assigned to them
from said first set of work partitions to perform said first operation;

assigning work partitions from said second set of work partitions to a second plurality
of entities, wherein said second plurality of entities are different entities than
said first plurality of entities; and

said second plurality of entities operating in parallel on work partitions assigned to
them from said second set of work partitions to perform said second
operation;

assigning work partitions from said third set of work partitions to said first plurality
of entities; and

said first plurality of entities operating in parallel on work partitions assigned to them
from said third set of work partitions to perform said third operation.

a'

86. The computer-readable medium of Claim 85 further comprising instructions for performing the following steps when a given entity in said first set of entities finishes performing a work partition from said first set of work partitions:

determining whether there are any unassigned work partitions from said first set of work partitions; and

if there are no unassigned work partitions from said first set of work partitions, then assigning the given entity a work partition selected from said third set of work partitions; and

if there are unassigned work partitions from said first set of work partitions, then assigning the given entity a work partition selected from said first set of work partitions.

87. The computer-readable medium of Claim 85 wherein the hierarchy includes odd levels and even levels, and the instructions further include instructions for performing the steps of assigning work partitions from odd levels to said first plurality of entities and work partitions from even levels to said second plurality of entities.

88. The computer-readable medium of Claim 85 wherein performing work partitions in said first set of work partitions causes said first set of entities produce output consumed by said second plurality of entities, and performing work partitions in said third set of work partitions causes said first set of entities to consume output produced by said second plurality of entities.

a' 89. A computer-readable medium carrying instructions for processing a query, the instructions including instructions for performing the steps of:
determining that, to execute said query, output from a plurality of producer operations
is to be supplied to a consumer operation;
wherein a first set of entities is responsible for executing a first producer operation of
said plurality of producer operations;
wherein a second set of entities is responsible for executing a second producer
operation of said plurality of producer operations;
wherein a third set of entities is responsible for executing said consumer operation;
during execution of said query, performing the steps of
determining whether any entity in said first set of entities produced output for
a particular entity in said third set of entities; and
if no entity in said first set of entities produced output for said particular entity
in said third set of entities, then communicating to at least one entity in
said second set of entities that is responsible for supplying output to
said particular entity that said at least one entity need not produce
output for said particular entity.

90. The computer-readable medium of Claim 89 wherein the step of determining whether
any entity in said first set of entities produced output for a particular entity in said third set of
entities is performed by
the particular entity monitoring whether it received any output from any entity in said
first set of entities;

a' the particular entity generating data that indicates whether it received any output from
any entity in said first set of entities; and
determining whether any entity in said first set of entities produced output for said
particular entity based on said data.

91. The computer-readable medium of Claim 89 further including instructions for
performing the steps of:

maintaining a bit vector, wherein each entity in said third set of entities corresponds
to a bit in the bit vector;

when all entities in said first set of entities that produce output for said particular
entity have completed their portion of said first producer operation, setting the
bit, in the bit vector, that corresponds to said particular entity if said particular
entity received no output from said first producer operation; and

wherein the step of determining whether any entity in said first set of entities
produced output for said particular entity includes the step of inspecting said
bit vector to determine whether any entity in said first set of entities produced
output for said particular entity.